First Project

Food2Vec

Context

OpenFoodFacts can be considerated as a wikipedia for food!
The goal of OpenFoodFacts is to share with everyone a maximum of informations on food products. It contains more than 2.5 millions products but maybe all products are not perfectly described... Mainly, for a product, we can find the list of ingredients, nutrition facts and food categories.

Expectations

A) Vectorize list of ingredients in order to learn a word2Vec model. So we assume that the context of an ingredient is defined by the other ingredients that often occur with it. Once the model learned, you have to create a map of ingredients! To do so, reduce your embeddings to 2 dimensions (with a PCA or a UMAP reduction).

Warning : you will be confront to two main problems, the mistakes on the vocabulary and the size of the dataset. You can resolve this issues selecting a subset of the data. Explain and justify the bias choosen for selecting your data (if you want to manage the entire dataset, go for it, but you will need some memory and some computation time).

B) Now you have vectors for ingredients, but how to use them to compare products?

Propose and implement a method allowing to compare automatically products.

With this method to evaluate Similarity between products, illustrate your approach on specific products: Select some products and show the most similar products found by your method.

C) Go forward and use your product similarity to achieve a map of products (like a Kmeans based on your product similarity for example).

Details

Your report must explain what technics/approachs you use, how you use them and the results obtained. You must deposit your report on DVO **before 2 december**.

If an approach don't work as planned you can show and explain (It will be very appreciate).
You can work in pairs of students. Your report must contain the names of students involved.
Your report must explain the logic of your approachs and results.
Your report must contain your link to your Colab Notebook. Your Deposit must contain a copy of your Notebook.
DataSet (approximately 7 Go): https://fr.openfoodfacts.org/data with different file formats available.

Please share your colab notebook with your group teacher:

aline.ellul@gmail.com

shikezhan@gmail.com

Christophe.rodrigues.bento@gmail.com

Good luck!